2019年6月5日　統計数理研究所　オープンハウス

マーク付き点過程の欠損データのテスト及び補完

Detection and replenishment of missing data for marked point processes

庄　建倉　　モデリング研究系　准教授

【Aim】

This project presents a fast approach for replenishing missing data in the record of a temporal point process with time separable marks, such as records of earthquakes and volcanic eruptions, in order to eliminate the underestimate of corresponding hazards caused by missing data.

【What is a point process】

Point process: random patterns of discreet events in space and/or time.

Marked point process: each event is assigned an attribute or size, namely a mark.

Conditional intensity: $\lambda(t, m \mid H_t)\, dt\, dm = \mathbf{E}\left[N\left([t, t+dt) \times (m, m+dm) \mid H_t\right]\right]$

Mark-separable marked point process: $\lambda(t, m \mid H_t) = \lambda_g(t \mid H_t) f(m)$

H_t: observation history up to time t.

【Bi-scale empirical transformation】

Bi-scale empirical transformation transforms a mark-separable point process

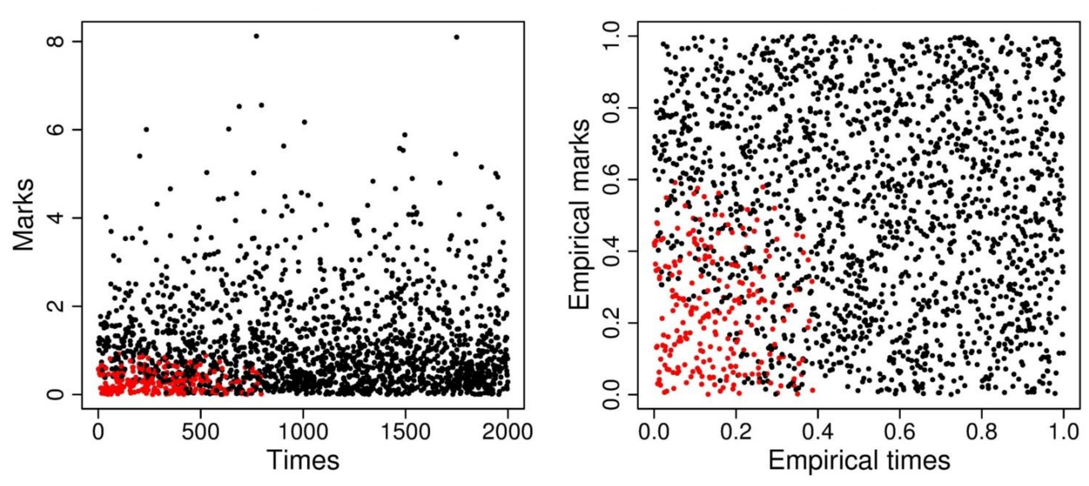
$N = \{(t_i, m_i): i = 1, \cdots N\}$ on $[0, T] \times \mathbf{M}$

into a homogeneous pattern on the unit square.

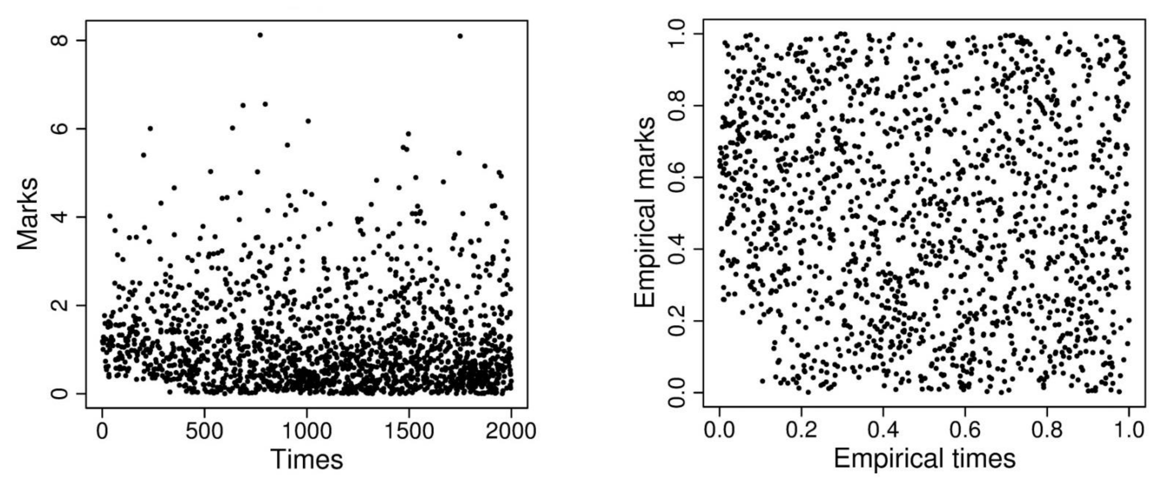
$$\Gamma: \quad [0, T] \times \mathbf{M} \to [0,1] \times [0,1]$$
$$(t, m) \to \left(\tilde{F}(t), \tilde{G}(m)\right)$$

$\tilde{F}(t), \tilde{G}(m)$: empirical distribution function of $\{t_i: i = 1, \cdots N\}$ and $\{m_i: i = 1, \cdots N\}$

Original complete process → *Homogeneous process*

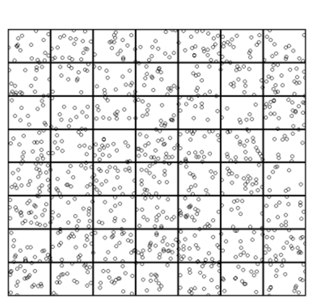


Original incomplete process → *Non-homogeneous process*



(*)

【Testing methods】



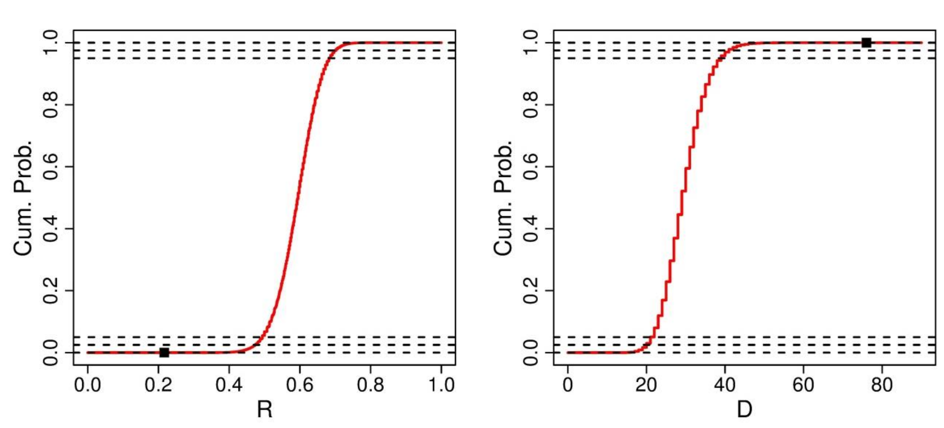
$$R = \frac{\min\{C_1, C_2, \cdots, C_L\}}{\max\{C_1, C_2, \cdots, C_L\}}$$

$$D = \max\{C_1, C_2, \cdots, C_L\} - \min\{C_1, C_2, \cdots, C_L\}$$

$$L = \#Row \times \#Column$$

C_i: number of events at each cells

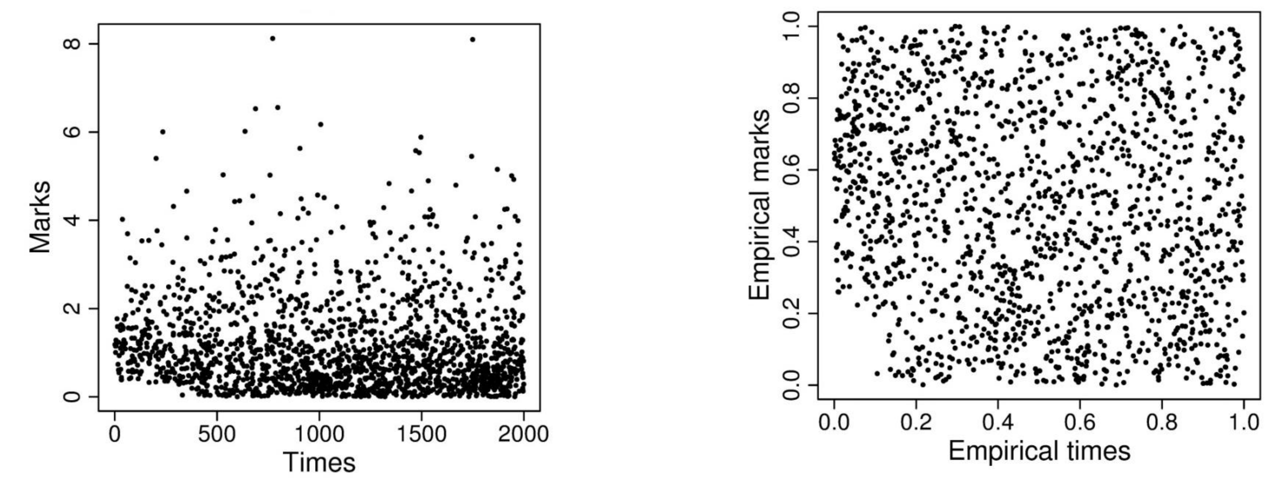
Testing results for (*)



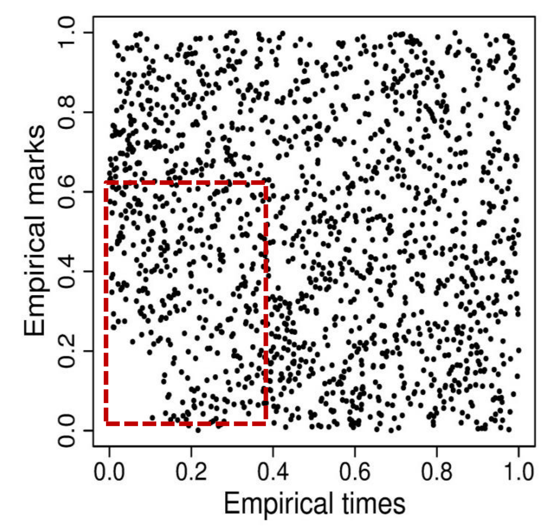
【Replenish Algorithm】

Step 1. Transform the process using the biscale empirical transformation.

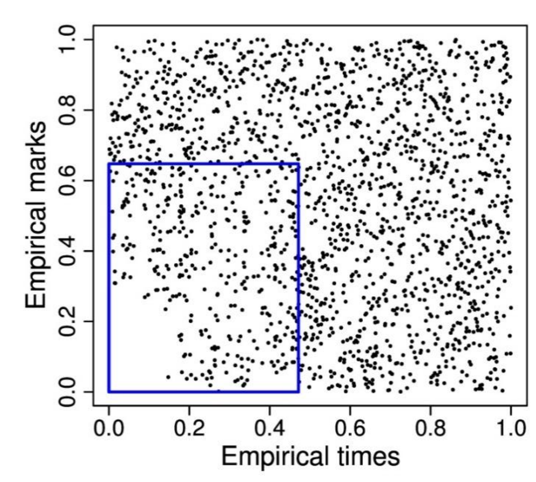
$t_i \to \quad \tau_i = \frac{i}{N} = \frac{\#\, of\, times < t_i}{N}$　　　$m_i \to \quad s_i = \frac{\sum_k \mathbf{1}(m_k < m_i)}{N} = \frac{\#\, of\, marks < m_i}{N}$



Step 2. Specify area S that contains the missing data in in the initial transformation domain.



Step 3. Calculate the missing area S^* in biscale transformation domain by solving



$$F_1^*(t) = \frac{\sum_{j=1}^{n} w_1(t_j, m_j, S)\mathbf{1}(t_j < t)}{\sum_{j=1}^{n} w_1(t_j, m_j, S)}$$

$$F_2^*(m) = \frac{\sum_{j=1}^{n} w_2(t_j, m_j, S)\mathbf{1}(m_j < m)}{\sum_{j=1}^{n} w_2(t_j, m_j, S)}$$

$$w_1(t, m, S) = \frac{\mathbf{1}\left((t, m) \notin S\right)}{\int_M \mathbf{1}\left((t, s) \notin S\right) dF_2^*(s)}$$

$$w_2(t, m, S) = \frac{\mathbf{1}\left((t, m) \notin S\right)}{\int_M \mathbf{1}\left((\tau, m) \notin S\right) dF_1^*(\tau)}$$

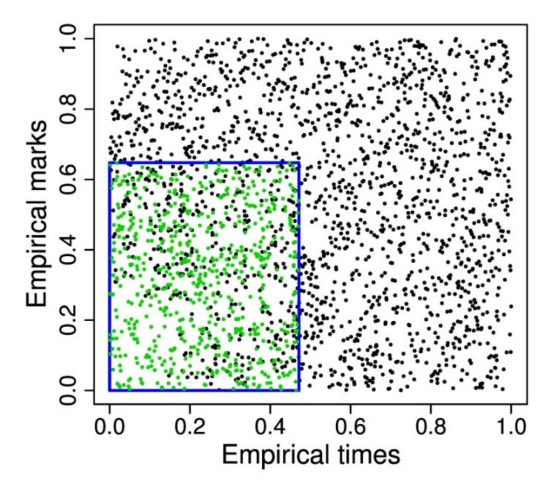
$$S^* = \{F_1^*(t), F_2^*(m): (t, m) \in S\}$$

Step 4. Generate data point in the missing area

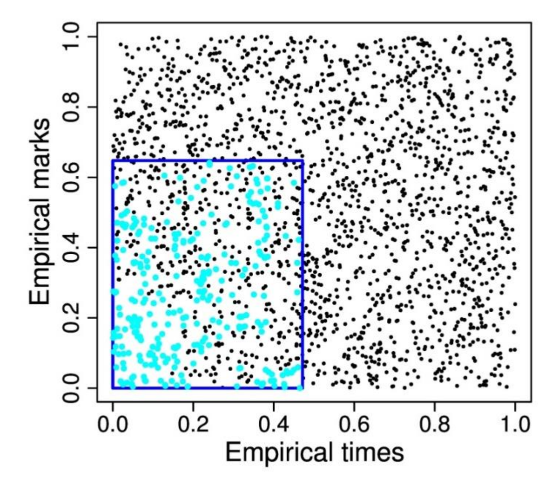
Generate events uniformly distributed in the missing region S^* (image of S)

\# events ~ $NB(k, 1 - |S^*|)$

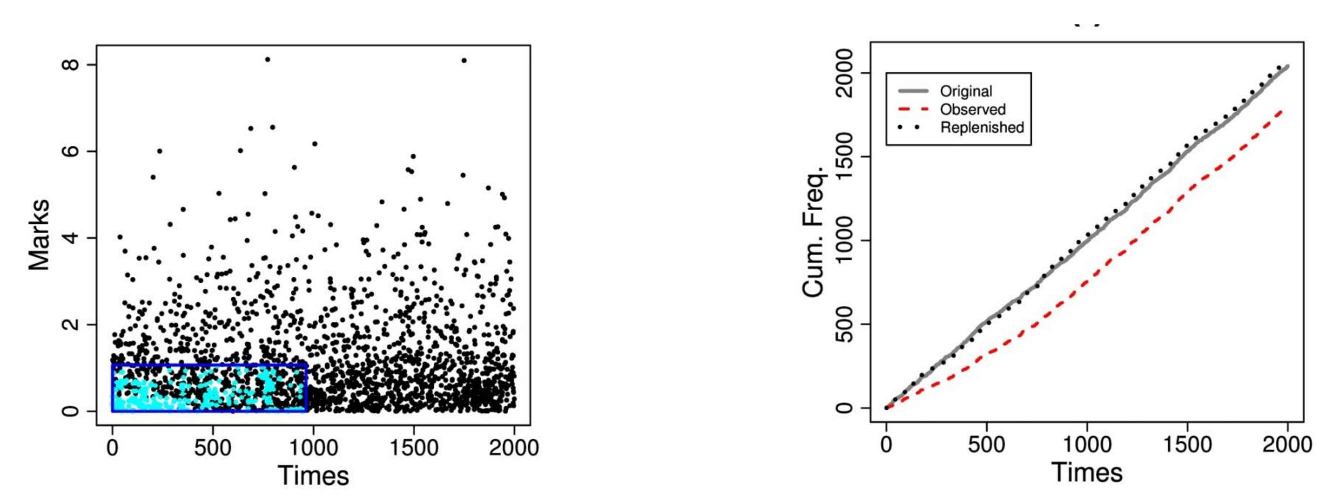
k: # of observed events outside of S^*



Step 5. Remove sequentially the closest simulated data point for each existing point in the missing area.

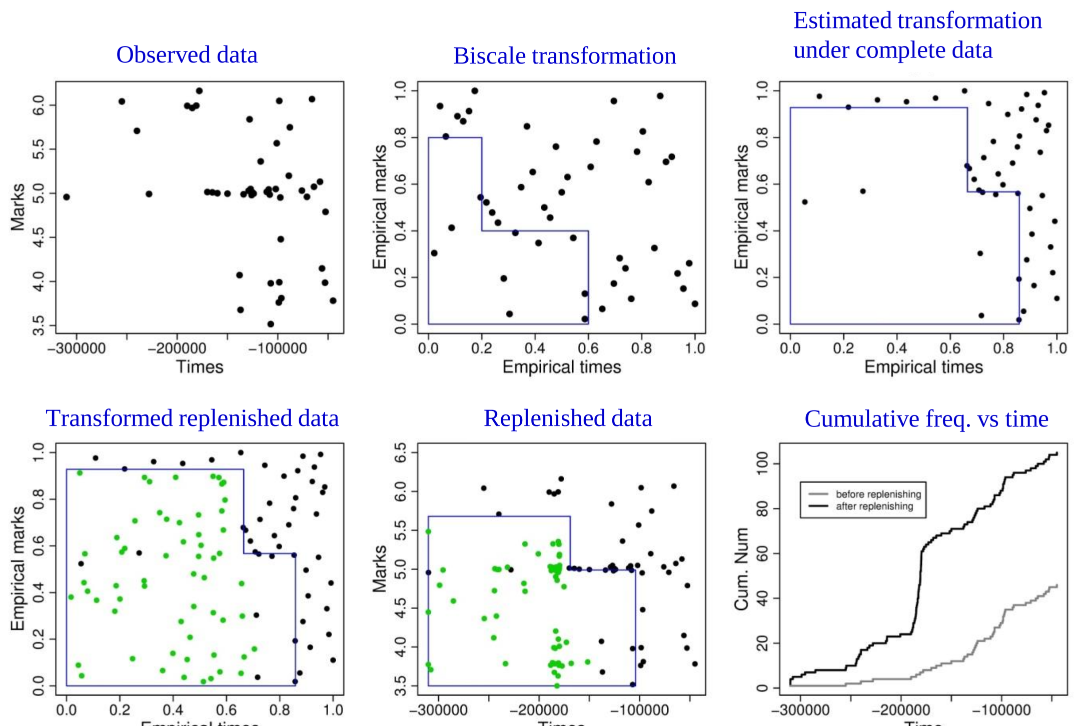


Step 5. Transform back all the events into the original domain



【Example 1: Eruptions at Hakone volcano】

Data source: (1) Smithsonian's Global Volcanism Program database (2) Large Magnitude Explosive Volcanic Eruptions database (LaMEVE database) (3) additional Japanese databases



【Example 2: Recent Kumamoto earthquake sequence data】

Data: Japan Meteorological Agency catalog
Time: 2016/4/1~2016/4/21
Mag.: 1.0+
Depth: < 100 km
Space: E128° - 133°
N30° -- 35°